Align Once to Explain: Feature Alignment for Scalable B-cosification of Foundational Vision Transformers

MAX PLANCK INSTITUTE FOR INFORMATICS

SIC Saarland Informatics Campus

Project Page

Raphael Maser*, Siddhartha Gairola*, Sukrut Rao, Bernt Schiele

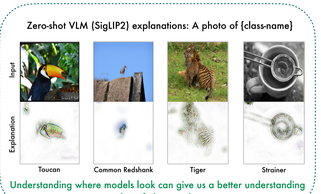
Max Planck Institute for Informatics,
Saarland Informatics Campus

Project Page

ellis European Laboratory for Learning and Intelligent Systems

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation



Understanding where models look can give us a better understanding about failure modes

Key Question

"How can we re-use existing foundation models to train interpretable equivalents?"

Why Does This Matter

Post-hoc methods are often unfaithful and compute-intensive. Inherently interpretable models like B-cos models provide faithful explanations by default.

B-cos Models

B-cos models have two important properties:

1. alignment pressure
2. linear summary

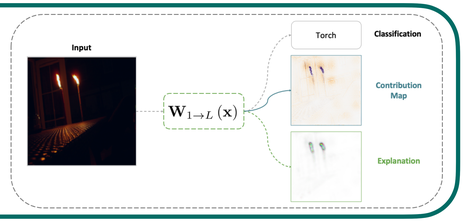
→ B-cos models are faithful

$$\text{Linear}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = \|\mathbf{w}\|\|\mathbf{x}\| \cdot \cos(\mathbf{x}, \mathbf{w})$$

$$\text{B-cos}(\mathbf{x}) = \|\hat{\mathbf{w}}\|\|\mathbf{x}\| \cdot |\cos(\mathbf{x}, \hat{\mathbf{w}})|^B \cdot \text{sign}\left(\cos(\mathbf{x}, \hat{\mathbf{w}})\right)$$

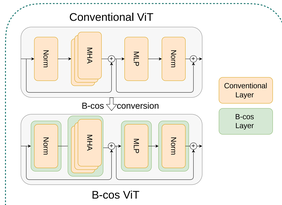


B-cos(x; w)



What do we do?

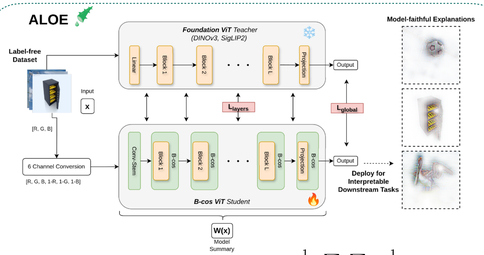
Step 1: Conversion



Replace:

1. linear projection by **convolutional stem**
2. linear layers by **B-cos layers**
3. **norms** by **bias-free** variant

Step 2: Distillation

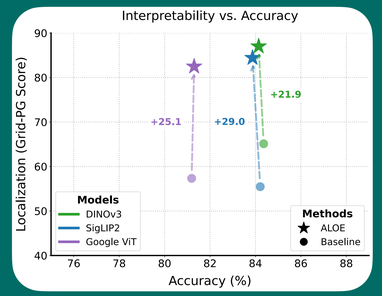


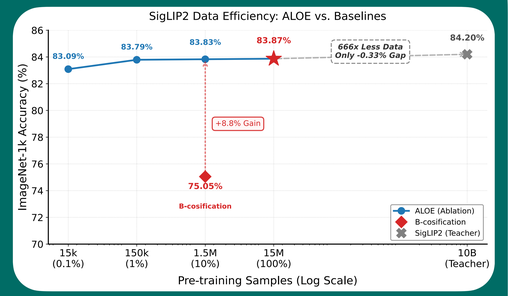
$$\mathcal{L} = \lambda_g \mathcal{L}_{\text{global}} + \lambda_l \mathcal{L}_{\text{layers}}$$

$$\mathcal{L}_{\text{global}} = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x}\in\mathcal{B}} \left(1 - \cos\left(E_S(\mathbf{x}),\, E_T(\mathbf{x})\right)\right)$$

$$\mathcal{L}_{\text{layers}} = \frac{1}{|\mathcal{B}|} \sum_{l\in\mathcal{L}_{\text{depth}}} \sum_{\mathbf{x}\in\mathcal{B}} \frac{1}{|\mathcal{T}^l_{\text{tok}}(\mathbf{x})|} \times \sum_{t\in\mathcal{T}^l_{\text{tok}}(\mathbf{x})} \left(1 - \cos\left(h^l_{S,t}(\mathbf{x}),\, h^l_{T,t}(\mathbf{x})\right)\right)$$

→ Minimize the mean cosine distance between layer-wise and global representations

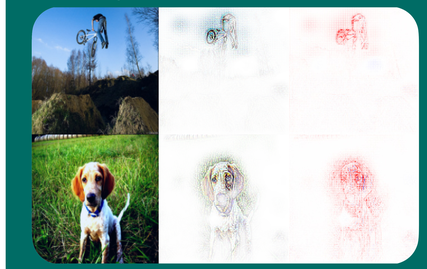




ALOE

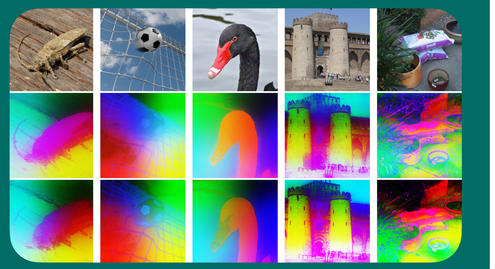
Efficiently Making Foundation Models Interpretable

Explanation Visualizations



Input ALOE ALOE *W*(**x**)

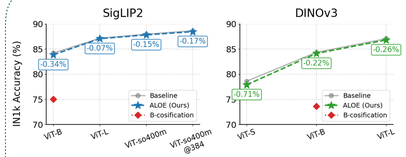
Feature PCA Visualizations



DINOv3 (row 2), ALOE (row 3)

What do we find?

Preserved Discriminative Capability

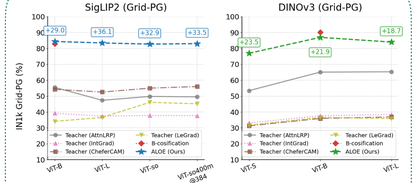


Architecture	Inh. Inter.	Zero-shot Acc.
Base architecture		
SigLIP2 [74] — ViT-B/16	✗	78.07
B-cosif. [4] — B-ViT-B/16	✓	61.01
ALOE (ours) — B-ViT-B/16	✓	**77.20**
Δ (*ALOE vs. B-cosif.*)		+16.19
Larger architectures		
SigLIP2 [74] — ViT-L/16	✗	82.28
SigLIP2 [74] — ViT-so400m/16	✗	82.56
SigLIP2 [74] — ViT-so400m/14@384px	✗	83.83
ALOE (ours) — B-ViT-L/16	✓	**81.97**
ALOE (ours) — B-ViT-so400m/16	✓	**82.39**
ALOE (ours) — B-ViT-so400m/16@432	✓	**83.13**

→ Same trend when evaluated on 10 fine- and coarse-grained datasets

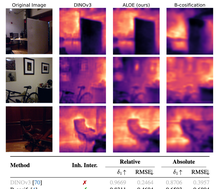
→ Similar zero-shot performance

Increase in Localization



→ Massively increased Grid-PG score

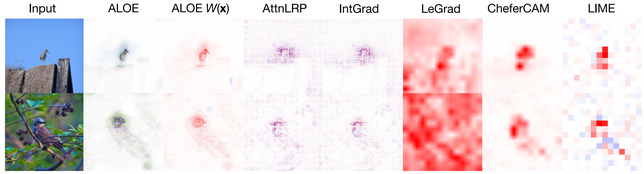
Dense Prediction



Method	Inh. Inter.	Relative $\delta_1\uparrow$	Relative RMSE↓	Absolute $\delta_1\uparrow$	Absolute RMSE↓
DINOv3 † [70]	✗	0.8907	0.3162	0.3705	0.6807
B-cosif. [4]	✓	0.8311	0.4604	0.6503	0.6804
ALOE (ours)	✓	**0.9416**	**0.2988**	**0.7850**	**0.4850**
Δ (*ALOE vs. B-cosif.*)		+0.1105	-0.1616	+0.1347	-0.1954

Depth estimation using Probe3D

Human Interpretable Explanations



→ Well localized explanations compared to post-hoc methods

References

1. Arya, Shreyash, et al. "B-cosification: Transforming Deep Neural Networks to be Inherently Interpretable." *NeurIPS* 2024
2. Böhle, Moritz, et al. "B-cos Networks: Alignment is All We Need for Interpretability." CVPR 2022
3. El Banani, Mohamed, et al. "Probing the 3D Awareness of Visual Foundation Models." CVPR 2024
4. Tschannen, Michael, et al. "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features." *arXiv* 2025
5. Siméoni, Oriane, et al. "DINOv3." *arXiv* 2025